Problem Set 1
Due Tuesday, October 20 at 11:59 AM (noon) via Canvas
ECON 31720, University of Chicago, Fall 2020

Assignments must be typeset in nicely formatted LaTeX. Programming assignments must use low-level commands, be commented clearly (not excessively), formatted nicely in 80 characters per column, etc. **You will be graded on the exposition of your written answers, the clarity of your code, and the interpretability and beauty of your tables and graphics.** The problem sets are individual assignments, but you may discuss them with your classmates. Submit the problem sets through Canvas in a single zip/tar/rar file. **Late problem sets will not be accepted under any circumstances.**

1. Suppose that we want to determine the causal effect of a binary variable $D \in \{0, 1\}$ on some outcome variable, Y. We also observe a scalar covariate, X. Let $Y(0)$ and $Y(1)$ denote the corresponding potential outcomes, and suppose that there are constant treatment effects, so that $Y(1)-Y(0) = \alpha$ is non-stochastic. Let β denote the population regression coefficient on D in a regression of Y on D and a constant. Let γ denote the population regression coefficient on D in a regression Y on D, a constant, and X. Assume throughout that both β and γ exist.

   (a) Is it true that $|\alpha - \gamma| \leq |\alpha - \beta|$? If so, prove it. If not, find a counterexample.

   (b) Suppose that D and X are uncorrelated. Does this change the answer to a)?

   (c) Suppose that X is uncorrelated with $Y(0)$ and $Y(1)$. Does this change the answer to a)?

   (d) Suppose that $\mathbb{E}[Y(0)|D = d, X = x] = \mathbb{E}[Y(0)|X = x]$. Does this change the answer to a)?

   (e) Suppose that $\mathbb{E}[Y(0)|X = x]$ is a linear function of x. Does this change the answer to a)?

2. (a) Suppose that we observe a scalar random variable Y. We know that Y is a mixture of two other random variables, X and Z, which we do not observe. That is,

   $$Y = WX + (1 - W)Z,$$

   where $W \in \{0, 1\}$ is a binary variable. Assume that both X and Z are independent of W. Let $G(y) \equiv \mathbb{P}[Y \leq y]$ denote the distribution function of Y, and let $F(y) \equiv \mathbb{P}[Z \leq y]$ denote the distribution function of Z. Show that

   $$\max\left\{\frac{G(y) - \pi}{1 - \pi}, 0\right\} \leq F(y) \leq \min\left\{\frac{G(y)}{1 - \pi}, 1\right\},$$

   for any y, where $\pi \equiv \mathbb{P}[W = 1]$ is a known probability. Show that these bounds are sharp (the best possible) for any fixed y when F can be any proper distribution function for a scalar random variable.

(b) Consider the same setting as in part a). Assume that Y, X and Z, are all continuously distributed, and let $G^{-1}(q)$ denote the qth quantile of Y. Show that

$$\mathbb{E}[Y|Y \leq G^{-1}(1-\pi)] \leq \mathbb{E}[Z] \leq \mathbb{E}[Y|Y \geq G^{-1}(\pi)],$$

and show that these bounds are also sharp. Explain how the bounds depend on π, and discuss the intuition.

(c) Suppose that we conducted an experiment to evaluate the impact of a job training program on wages. Let $D \in \{0,1\}$ denote participation in the job training program. Let $Y(0), Y(1)$ be potential outcomes that denote wages one year after the program depending on an individual's participation in the program, and let $Y^\star = DY(1) + (1-D)Y(0)$.

Unfortunately, we do not observe $Y^\star$, because not everyone is employed one year after the experiment. Let $S \in \{0,1\}$ denote employment, and let $S(0)$ and $S(1)$ denote potential employment depending on an individual's participation in the program, so that $S = DS(1) + (1-D)S(0)$. Then we observe $Y = Y^\star$ if $S = 1$, but we do not observe any wage data if $S = 0$. We do observe both D and S for everyone.

Assume that the experiment was conducted perfectly, so that D is independent of $(S(0), S(1), Y(0), Y(1))$. Also, assume that the job training program made everyone more likely to be employed, so that $\mathbb{P}[S(1) \geq S(0)] = 1$, and that we observe wages for at least some people in both treated and control arms, so that $\mathbb{P}[S = 1|D = d] > 0$ for $d = 0, 1$. Let $\mu = \mathbb{E}[Y(1) - Y(0)|S(0) = 1, S(1) = 1]$ denote the average treatment effect for those who would be employed one year after the program, even if they hadn't taken it. Show that

$$\begin{aligned}\mathbb{E}[Y|D=1, S=1, Y \leq \bar{G}^{-1}(1-\pi)] - \mathbb{E}[Y|D=0, S=1]\\ \leq \mu \leq \mathbb{E}[Y|D=1, S=1, Y \geq \bar{G}^{-1}(\pi)] - \mathbb{E}[Y|D=0, S=1],\end{aligned}$$

where $\bar{G}^{-1}(q)$ is the qth quantile of the distribution of $Y|D = 1, S = 1$, and

$$\pi \equiv \frac{\mathbb{P}[S=1|D=1] - \mathbb{P}[S=1|D=0]}{\mathbb{P}[S=1|D=1]}.$$

Explain why these bounds are sharp.

3. Suppose that we observe a discretely distributed treatment variable D that takes values in some finite set $\mathcal{D}$, an outcome Y and some covariates X. Let $\{Y(d) : d \in \mathcal{D}\}$ denote the potential outcomes for Y. Assume that the distribution of $Y(d)$ conditional on $D = d, X = x$ is the same as the distribution of $Y(d)$ conditional on $X = x$ for all d and x. For any $d \in \mathcal{D}$, let $p(d, x) \equiv \mathbb{P}[D = d|X = x]$, and let $P_d \equiv p(d, X)$. Show that

$$\mathbb{E}[Y|D = d, P_d = p] = \mathbb{E}[Y(d)|P_d = p]$$

for any jointly supported $d \in \mathcal{D}$ and $p \in (0,1)$. Explain what this result shows and why it is significant for empirical practice. Then, let $P \equiv p(D, X)$ and show that

$$\mathbb{E}[Y(d)] = \mathbb{E}\left[\frac{Y\mathbb{1}[D=d]}{P}\right],$$

where we assume that $P > 0$ with probability 1.

4. Suppose that

$$Y = \sin(2X) + 2\exp(-16X^2) + U,$$

where U is normally distributed with mean 0, standard deviation .3, and X is uniformly distributed over $[-2, 2]$. We want to estimate $m(x) \equiv \mathbb{E}[Y|X = x]$ nonparametrically. Conduct a Monte Carlo simulation that demonstrates the bias-variance trade-off in the context of nonparametric regression using the following methods:

   - Local constant (kernel) regression.
   - Local linear regression.
   - A sieve approximation using the standard polynomial basis (e.g. $1, x, x^2, x^3, \ldots$).
   - The nearest neighbors estimator.
   - A sieve approximation using the Bernstein polynomial basis. A Bernstein polynomial of degree K is given by

   $$B(z) \equiv \sum_{k=0}^{K} \theta_k b_k^K(z) \quad \text{where} \quad b_k^K(z) \equiv \binom{K}{k} z^k (1-z)^{K-k},$$

   for coefficients $\{\theta_k\}_{k=0}^K$, where $z \in [0, 1]$. Note that you construct a Bernstein polynomial on a compact domain other than $[0, 1]$ (such as $[-2, 2]$) by setting $z = (x - (-2)/(2 - (-2))$.
   - A sieve approximation using linear splines represented in the truncated power basis, that is, a function of the form

   $$f(x) = \theta_0 + \theta_1 x + \sum_{k=2}^{K+1} \theta_k \mathbb{1}[x \geq r_{k-1}](x - r_{k-1}),$$

   for coefficients $\{\theta_k\}_{k=0}^K$, where $r_1 < r_2 < \ldots < r_K$ are known "knots." A common way to set the knots is to take r_k to be the $k/(K + 1)$th quantile of X.

   Report your results graphically by plotting the mean and standard deviation of your estimates (as a function of x) across simulation draws for two different values of the tuning parameter. I recommend two side-by-side subplots plots per method (one per tuning parameter), but plotting the results with both tuning parameters on the same plot might be feasible too.

   Your work will be graded on how compelling your simulation is, including the clarity of your exposition of the results.

5. This question is about "Persecution Perpetuated: The Medieval Origins of Anti-Semitic Violence in Nazi Germany" by Nico Voigtländer and Hans-Joachim Voth, published in *The Quarterly Journal of Economics* in 2012. The paper, as well as the data and code used in the paper, are available on Canvas. Read as much of the paper as you need to answer the following questions. (No need to read the whole thing; many parts will be irrelevant.)

(a) The authors seem to argue that conditioning on covariates is important. Given their argument, why is the set of covariates that they use in (1) a bit odd? *(Hint: Read footnote 37.)*

(b) Replicate column (1) of Table VI. This may be difficult, especially panels B and C, but give it a good shot. Remember to read the table notes carefully and to consult the authors' Stata code.

(c) Implement propensity score matching estimators of both the ATE and ATT, using the same covariates as the authors do. I leave the specifics up to you, but you might consider nearest neighbor matching on the propensity score, and/or a blocking approach. Compare your estimates to the authors' estimates. You may use the bootstrap to compute standard errors.
Note: Optimization packages are not considered high-level commands for the purpose of this class, since they are not statistical in nature. You may (and should) use one to optimize a likelihood, e.g. for a logit estimator.